SRA Submission Guidelines

National Center for Biotechnology Information (NCBI)

National Library of Medicine

22 Mar 2010 Version 1.1 Draft B

Contents

1 Overview

1.1 Scope

The Sequence Read Archive (SRA) at NCBI accepts primary sequencing data from so-called "next generation" sequencing platforms, including Roche 454®, Illumina Illumina®, Applied Biosystems SOLiD®, Helicos Biosciences HeliScope®, CompleteGenomics® and others.

Sequencing data should be submitted to the SRA rather than the regular Trace Archive. The Trace Archive is intended as the repository of sequencing data from gel/capillary platforms (Applied Biosystems 370® and 3730® , Megabace, and Licor sequencers).

1.2 Related Documents

- NCBI Sequence Read Archive
- SRA XML Schema Documents
- Sequence Read Format
- Aspera Download Guide
- Submission Quick Start Guide

1.3 Revision History

Release	Notes
11 Sep 2009 Version 1.0 Draft A	Production version
25 June 2008 Version 0.8 Draft B	Last provisional version

1.4 Links and Contacts

http://www.ncbi.nlm.nih.gov/Traces/sra	SRA Home Page at NCBI
trace@ncbi.nlm.nih.gov	The Trace Archives mailing list for inquiries
Trace Help Desk	Specific problems or issues regarding trace submissions (web form)
sra@ncbi.nlm.nih.gov	Aspera technical support
http://www.asperasoft.com	Asperasoft home
http://en.wikipedia.org/wiki/EXT3	ext3 filesystem

http://en.wikipedia.org/wiki/NTFS	NTFS filesystem
http://en.wikipedia.org/wiki/Linear_Tape-Open	Linear Tape Open (LTO) tape formats
http://en.wikipedia.org/wiki/Ftp	File transfer protocol (ftp)
http://en.wikipedia.org/wiki/Http	Hypertext transfer protocol (http)
http://en.wikipedia.org/wiki/USB	Universal Serial Bus (USB) connections
http://en.wikipedia.org/wiki/User_Datagram_Protocol	User Datagram Protocol (UDP)

2 Terms of Usage

2.1 Permanence

Accessions issued by the SRA are always maintained and never reused. If a desired record has been withdrawn, then a message to this effect will be displayed to anyone who tries to access it. If a record has been superseded by a successor record, this fact will be presented to anyone trying to access it. Only in rare cases where the record needs to be expunged from the archive will a user not be able to access it.

2.2 Authentication

Submissions are managed through secure channels. These channels include MyNCBI, NIH level login through CIT, and ftp accounts secured by passwords. We will correspond with submitters via email about submission and curation issues, but we do not exchange data by email.

Individuals may obtain a MyNCBI account for their transactions. Please do not reuse someone else's MyNCBI account. Center accounts are provided for the convenience of automated pipelines. The authentication information for such an account must be maintained securely by the Center. Accounts may be disabled or withdrawn after a long period of disuse in order to comply with NCBI security requirements.

2.3 Limitations

The Sequence Read Archive at NCBI is a public resource and the decision whether to submit data to this resource is the responsibility of the submitter. Prospective submitters should be aware of the following issues:

Never submit data without the permission of the **principal investigator**.

Most **human data** gathered from research subjects are under strict privacy controls and/or usage restrictions and must be handled with protections as determined by the research institution's Institutional Review Board (IRB), the funding agencies, and the

laws of the United States or the submitter's home country. The dbGaP resource at NCBI may be a more appropriate broker for human sequencing data requiring controlled access due to these considerations.

Data submitted as part of a journal manuscript may have a **publication embargo** placed on it by the journal editors. The submitter can place a "hold until publish" restriction on the submission to the SRA as part of the submissions process.

Data that might relate to **patents** and **intellectual property** may be submitted to NCBI, but the submitter is responsible for ensuring that procedures and policies of his/her institution or company are observed.

Some **environmental data** gathered in the territory of certain countries, including territorial waters, may have sovereign legal restrictions on their use. NCBI cannot accept such data since NCBI is not able to enforce any usage restrictions..

Submitters must ensure that data obtained as part of a criminal investigation is free of any judicial restrictions on its use.

Submitters are responsible for obtaining any necessary permissions from the collecting institution for **forensic and paleontological data.**

2.4 Modification

NCBI allows submitters to modify their records. Such requests must be formally entered using the SRA submission mechanisms. Informal requests by email will not be accepted. Only the center or individual that created the record can change it. Please write NCBI if you have changed affiliations and wish to update old records. This may require agreement from the original institution.

2.5 Curation

From time to time records deposited at the SRA must be updated with changes needed in order that the data continue to conform with the data model for the archive, to update data as it changes (for example finalizing publication information), to change data that are clearly wrong (for example correcting external references to other data or resources), and to add additional relevant metadata as they become available. NCBI will contact the submission owner on a best effort basis. The submission owner should maintain up to date contact information with NCBI to receive word of such changes.

Actual instrument data are never changed by NCBI. Only the submitter can make such modifications.

2.6 Availability

While NCBI tries to maintain maximum uptime of its servers on a 24x7 basis, no guarantee of availability is offered to users. Submissions that are interrupted by downtime may have to be restarted by the user.

Technical assistance is available on a limited basis during business hours USA Eastern Time. There is no guarantee for level of service regarding manual assistance.

3 Data Model

The SRA data model is discussed in detail elsewhere, but here is a brief overview.

The SRA tracks the following five objects:

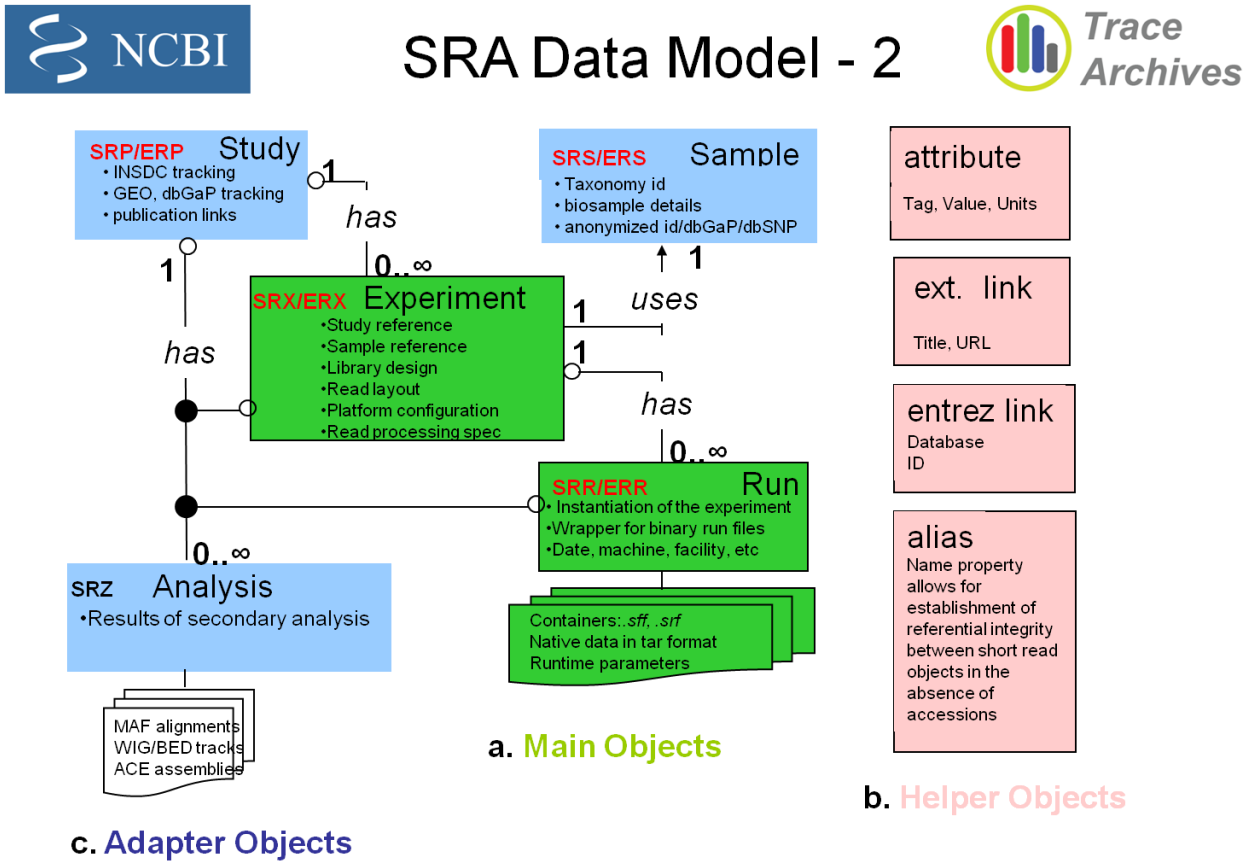
Study – Identifies the sequencing study or project and contains multiple experiments.

Sample – Identifies the organism, isolate, or individual being sequenced.

Experiment – Specifies the sample, sequencing protocol, sequencing platform, and data processing that will result one or more runs.

Run – Identifies run data files, the experiment they are contained in, and any runtime parameters gathered from the sequencing instrument.

Analysis – Packages data associated with short read objects that are intended for downstream usage or that otherwise needs an archival home. Examples include assemblies, alignments, spreadsheets, QC reports, and read lists.



In addition, all details concerning submissions are contained in a separate document called **Submission**, which contains center specific submitting information, contacts, actions for the archive, and a file manifest.

Objects can be archived in the SRA at different points in time. Multiple submissions documents can be submitted. For example, study, sample, and experiment objects can be created at an early stage, with run data being submitted as the data are produced.

All SRA objects that are being created with XML files can be referenced by an alias. This is even true after they have received an accession. The namespace that the alias must be unique in is that of the submitting center.

4 Obtaining NCBI Accounts Needed for Submission

4.1 Establish a NCBI Identity

Before interacting with NCBI, please obtain a personal identity. This will allow you to make submissions, track results, change records now or later, and hold or release records. There are three kinds of identity each of which is sufficient to do business with NCBI:

NCBI PDA – NCBI-created and managed account for primary data submitters. If you belong to a submitting Center and will play a role in monitoring and maintaining primary data submissions, please identify this fact through your account profile and also email NCBI with this information.

NIH – For any NIH personnel who has credentials managed through CIT and can use their NIH identity to login

4.2 Establish a Center Name

Certain submitters might find it necessary to establish a more formal relationship for their sequencing center or lab. These are typically large groups that are planning to use an automated system to create submissions and transfer their files.

	Individuals	**Centers**
Tracking	Interactive tool only	XML telemetry available
Submissions	Low frequency	Often > 10 per year
Contact Information	PDA account	Permanent contacts required
Time to Submit	Immediate	Requires setup
Size of Files	Files usually < 4Gb (due to FTP limits)	Any file size
Users Able to Update	Submitter only	Any account linked to center

Maintenance	Interactive tool only	Interactive or XML updates
Status Notices	No notices	Notice for: Updates outages
Uploads	FTP only	FTP or Aspera

If your lab, center, or group has submitted in past, there might already be a center established. Please check here for your center or lab. ftp://ftp.ncbi.nlm.nih.gov/sra/etc/centers.tab Please contact sra@ncbi.nlm.nih.gov to be added to an established Center's user list or to provide information for creating a new Center.

To create a new Center, please provide the following information:

1. suggested center abbreviation (8 char max)
2. center name (full)
3. center URL
4. center mailing address (including country and postcode)
5. phone number (main phone for center or lab)
6. contact person (someone likely to remain at the location for an extended time)
7. contact email (ideally a service account monitored by several people)

Please read section 5.3 Transmitting Data to NCBI to select a method of data transfer for your center.

5 Submitting Data

5.1 Understanding Submission Modes

5.1.1 High Throughput Submissions

High-volume submissions should be uploaded to the ftp directory for your center. To do this,

ftp: ftp-trace.ncbi.nlm.nih.gov

login: myaccount_trc

passwd: !jXYZZ3@ce

> cd short_read

> put myfiles.tgz

> quit

You should double check that the file size that you posted agrees with the original file.

You cannot delete files once they are posted. Please write to sra@ncbi.nlm.nih.gov to request deletions.

5.1.2 Individual Submissions

Individual submitters may submit files through anonymous ftp to NCBI. This account is shared by all such submitters, but the files once written are not readable even by the submitter. In this case the center name is "Individual" and submissions are not tracked by institution. Please ensure that contact information written into the submission documents will allow NCBI to contact you to confirm receipt of the files and to communicate any problems.

To submit individually, please

- Create a NCBI PDA account
- Write trace@ncbi.nlm.nih.gov to request the ftp address of the current anonymous ftp box.
- Put the file to the anonymous ftp box, for example *mysubmission.tar.gz*

The Individual Submissions channel is intended for small submissions or uploads of test submissions when the submitter does not yet have a private account.

5.1.3 Interactive Submissions

You can use a web tool to create and manage your SRA submissions. You will need a NCBI PDA account. Please visit
http://www.ncbi.nlm.nih.gov/Traces/sra_sub/sub.cgi?&m=submissions&s=default

You must login with, or create a new NCBI PDA account in order to proceed. MyNCBI accounts do not work.

5.2 Packaging Data for Submission

NCBI provides a flexible environment for detecting and processing submissions

5.2.1 Data for Interactive Submissions

You will need to transmit through ftp your run data files (SFF, SRF, etc). Please do not compress these files, it only adds delays to processing and archival of the data.

5.2.2 Bulk Submissions

For bulk submissions where XML documents describing the submission and metadata accompany the run data, please follow these guidelines:

Always include a submission.xml file with your submission

Please do not send bare xml files, always package them in a tar file.

Please send run data files separately from XML, unless all files can be easily contained within one tar file.

Please do not compress files. This only adds delays to the processing of the submission.

5.3 Transmitting Data to NCBI

You will have to somehow transmit your run data files to NCBI. This cannot be done through the interactive submission tool. Run data files (SFF, SRF, etc) can be quite large. It is NOT recommended to compress these, as they are already compressed to some degree and decompression adds to the archival processing time and will delay submission and release of the data.

5.3.1 ftp

The ftp service provided to established centers has long been the normal method for transferring trace data with NCBI. Users are recommended to switch to Aspera client for downloads, and to use *ascp* copy program for uploads.

5.3.1.1 Limitations using ftp

Traditionally NCBI has relied on *ftp* as the means for transferring large files. Bandwidth on transfers is typically 100 Mbps with less on international transfers. NCBI does not impose an upper limit on ftp transfer size. However, maximum file sizes above 10 GB may fail due to limitations elsewhere in the path from center to NCBI.

5.3.1.2 Bulk Submissions via ftp

High-volume submissions should be uploaded to the ftp directory for your center, which is provided with the secure ftp account provisioned to your center.

For example, a user working for the *mycentre* center will deposit short read data into the short read directory of the ftp account's login directory as follows:

ftp: ftp-trace.ncbi.nlm.nih.gov

login: mycentre_trc

passwd: !jXYZZ3@ce

> cd short_read

> put myfiles.tar.gz

> quit

You should double check that the file size that you posted agrees with the original file.

5.3.1.3 Individual Submissions via ftp

The NCBI Trace Archives maintains a private ftp address. Please write to mailto: sra@ncbi.nlm.nih.gov for the current address, which contains both the ftp address and login string.

The unix/linux shell command will look something like:

ftp `ftp://sra:0!8e5frRy!@ftp-private.ncbi.nih.gov/`

Please observe the following rules when using this submission method:

Maximum 1 Gigabyte file size

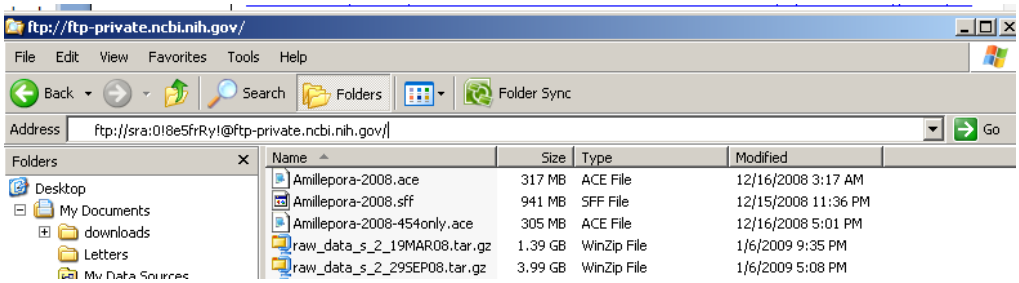
Maximum 10 file limit per submission

Choose a unique filename that also will be easy for you to identify

This directory has special access rules. You can stat the directory (list the files), but you cannot read any file (or download a file), and Once deposited, the file cannot be overwritten. The files are removed as soon as processed, or if they have remained too long on the server. It is your responsibility to complete the submission transaction in a reasonable amount of time so that the files you have deposited through this channel can be processed by the submission system.

5.3.1.4 ftp from Windows

It is possible to upload to NCBI ftp sites from Windows. Use Windows Explorer to access the individual ftp address as follows. Then simply drag and drop the submission files from your source directory into the destination directory that the Explorer tool has opened.



You can also login using your center account, and utilize Windows Explorer to navigate and upload.

5.3.1.5 Troubleshooting ftp

If you are having trouble with your ftp connection to NCBI, try

1. Setting passive mode rather than active mode
2. Ask your sysadmin to increase ftp buffer size to 32 MB
3. Try another host, or another platform (Windows instead of Unix)
4. Try using the unix *split* utility to split up the transfer file into smaller pieces. Be sure to provide reassembly instructions and checksum of the reassembled file to verify rebuilding the original file on our end.
5. Try another ftp client software:

   *ncftp (*http://www.NcFTP.com*)*

Windows *filezilla* (http://filezilla.sourceforge.net)

If you still have trouble, please write us with the following details:

1. time of transfer (GMT or local time)
2. IP address of ftp client (the system you are doing ftp from)
3. version of unix software (uname –a, or cat /proc/version)
4. ftp account used
5. specific error messages (connection closed, etc)

5.3.2 Disk and Tape

Archive users can also request or submit data on disk or tape. The following are requested:

- LTO4 (we can also read LTO3 and LTO2)
- HDD with USB2.0 or FireWire interface enclosure with WinNT (FAT32) partition type, so any Windows or Linux computer can read them.
- NTFS, Ext3, or other large format drives. Please ensure they are delivered with an enclosure. We prefer FireWire interface.

To get the submitted media returned, please plan on providing a waybill for shipping. If you are requesting a download by disk or tape, please send us the media first, along with a waybill for return shipping.

Please use the following shipping address:

Martin Shumway, Staff Scientist
DHHS/NIH/NLM/NCBI
45 Center Drive
Bldg. 45/Room 6A N 24
MSC 6510
Bethesda, MD 20892
shumwaym@ncbi.nlm.nih.gov
tel: 301.402.4041
fax: 301.402.9651

5.3.3 Aspera

5.3.3.1 The fasp Protocol

The FASP protocol from Aspera (www.asperasoft.com) uses UDP, eliminating the latency issues seen with TCP, and provides bandwidth up to 1 Gbps to transfer data. It has a restart capability if data transfer is interrupted midstream and is well behaved, so if there is other data traffic on your network connections, it will back off in order to avoid starving other protocols. We have seen effective throughput up to 600 Mbps to a single site.

NCBI is implementing Aspera for two use cases, occasional users and those who download files for direct use (Aspera Connect), and bulk users who will be uploading or downloading large amounts of data (ascp)

5.3.3.2 Aspera Connect

Aspera Connect is software that allows download and upload via a web plugin for popular browsers for machines running Linux, Windows, and Macs and a command line tool that allows scripted data transfer. The software client is free for NCBI site users for the purpose of exchanging data with NCBI.

Download and install AsperaConnect software:

http://www.asperasoft.com/downloads

Select the Connect product for your browser and platform. By default, the plugin configuration is less than optimal. To change, right click the Aspera icon from the system tray. Select ‘networks’ and update the connection speed (e.g. 622Mbps).

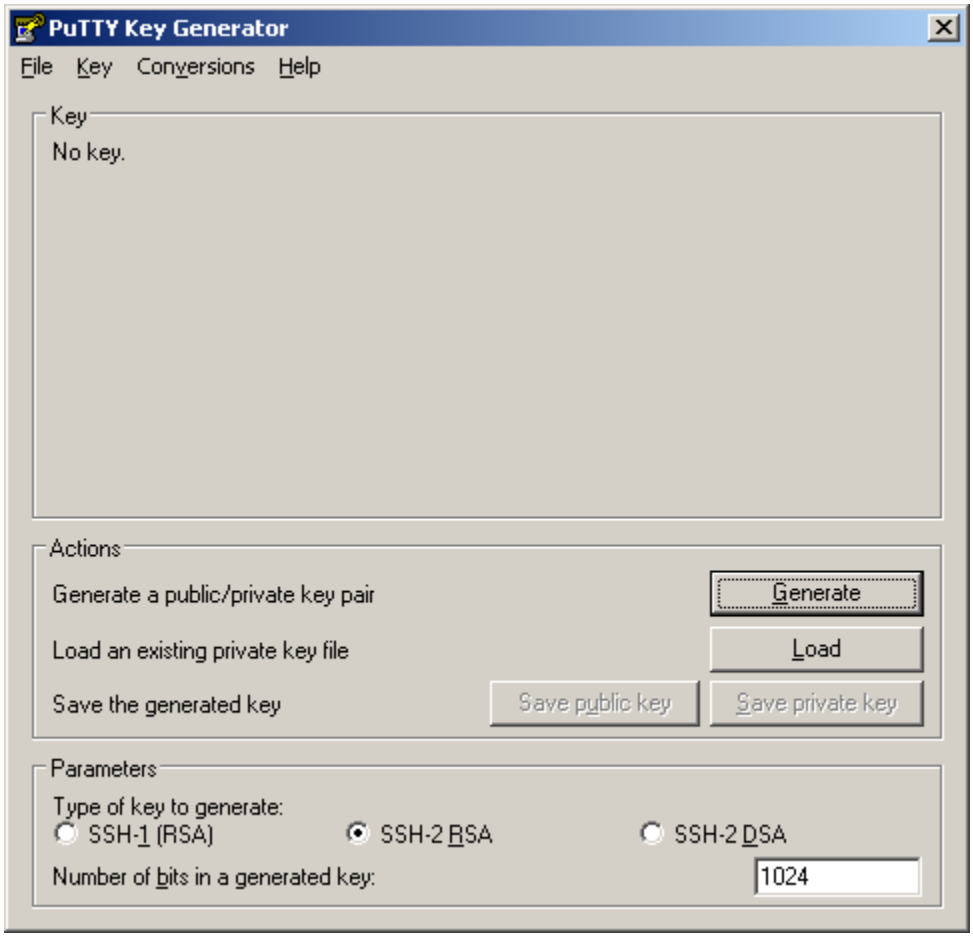
5.3.3.3 Setting Up Aspera for Bulk Transfers

Tell NCBI you are preparing to set up a link, and we will provide a login account. There will be one account per center. Please set up a Center identity for your institution or lab if you do not already have one.

Your local firewall must permit UDP data transfer on port 33001-33009 in both directions to allow the fasp traffic to pass and must allow ssh traffic outbound to NCBI.

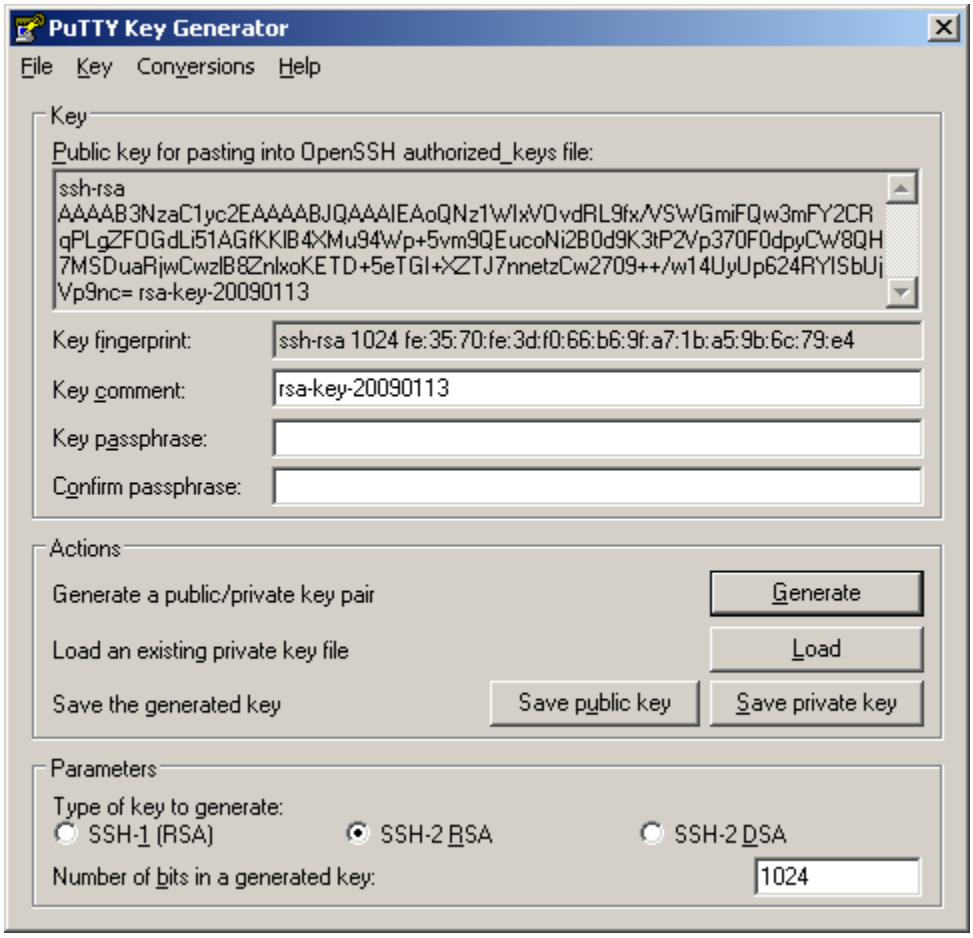
Download puttygen: http://the.earth.li/~sgtatham/putty/latest/x86/puttygen.exe

Run puttygen.exe to create ssh key:



Make sure that SSH-2 RSA Parameter option is selected, and that the "Number of bits in a generated key" be set to 1024. Then press "Generate" (moving mouse to generate key).

Generating a key will result in something like this:



Click “Save Private Key” to retain the private key. NOTE – leave “Key passphrase” and “Confirm passphrase” empty (otherwise, you will be prompted to enter the passphrase whenever you do an Aspera transaction).

Copy the text from the “Public Key for pasting into OpenSSH authorized_keys file” text box:

```
ssh-rsa
AAAAB3NzaC1yc2EAAAABJQAAAIEAoQNz1WIxVOvdRL9fx/VSWGmiFQw3mFY
2CRqPLgZFOGdLi51AGfKKlB4XMu94Wp+5vm9QEucoNi2B0d9K3tP2Vp370F
0dpyCW8QH7MSDuaRjwCwzlB8ZnlxoKETD+5eTGI+XZTJ7nnetzCw2709++/
w14UyUp624RYISbUjVp9nc= rsa-key-20090113
```

In order that a submission center is able to access (i.e. transfer and receive files from NCBI using Aspera Connect), this public ssh key must be provided to NCBI. This key should be emailed to: sra@ncbi.nlm.nih.gov with subject line “Aspera connect authorization request”.

SSH keys are used for establishing secure connections to remote computers

5.3.3.4 Using ascp for Bulk Transfers

The command line program *ascp* is a utility delivered along with the AsperaConnect product.

You can run the *ascp* program with the following parameter settings:

–Q (for adaptive flow control)

–l (maximum bandwidth of request, try 200M and go up from there)

–m (minimum bandwidth of request, try 0)

–r recursive copy

–T no encryption (speeds up transfers). Connection remains secure however data being transferred is not.

–i <private key file>

Try experimental transfers starting at 100 Mbps and working up to 400-500 Mbps. Select the bandwidth setting that gives good performance with unattended operation. Copy the file to:

ascp -i <private key file> -QTr <file(s) to transfer> -l100M asp- @upload.ncbi.nlm.nih.gov:test/

where

<private key file>	::= fully qualified path & file name where the generated private key was saved.
<files(s) to transfer>	::= names of files to transfer (including path)
	::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence.
100M	::= tunable mbit/sec bandwidth

The *ascp* command on Microsoft Windows is located by default in `c:\program files\aspera\Aspera Connect\bin\ascp`

The *ascp* program on Mac in located at *aspera/bin/ascp*

The *ascp* program on Linux is located at *<install directory>/bin/ascp*

It is possible to run *ascp* in an autonomous, unattended manner that does not require repeated login. Please send us the public key of a SSH key pair and we will add it to our authentication system.

5.3.3.5 Pushing Data with ascp

Use the command line utility *ascp* to copy files directly to a remote host:

ascp -i <private key file> -QTr <file(s) to transfer> -l300M \ asp- @upload.ncbi.nlm.nih.gov:incoming/

where

<private key file>	::= fully qualified path & file name where the generated private key was saved.
<files(s) to transfer>	::= names of files to transfer (including path)
	::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence.
300M	::= tunable mbit/sec bandwidth

5.3.3.6 Administering Remote Files

Do **not** delete files in order to "make space". The SRA is responsible for maintaining adequate space by removing files that have already been processed. If files are not being deleted it is likely because of a backlog in the SRA.

If a submission file needs to be replaced, wait until you have a replacement and then overwrite the file (do not delete it). Please DO replace zero length files or files that have been truncated. If a "junk" file has been transmitted by mistake, it can be removed.

NOTE - files that have not been attached to any submission may be deleted after a certain amount of time. It is recommended that you consult with the SRA Administrators for the current expiration policy.

You may establish a secure connection to the SRA by using putty.exe along with your private ssh key. For example:

putty.exe –i <private key file> asp- @upload.ncbi.nlm.nih.gov

where

<private key file>	::= fully qualified path & file name where the generated private key was saved.
	::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence

Once connected, you may use the '*ls*' command to view the directory. You will not be able to change directories (e.g., use of the '*cd*' command is disabled). Valid *ls* commands include:

ls -l test #lists the content oftest subdir in long format
ls -l incomimg #lists the content of incoming subdir in long format
ls -l #lists the content of home directory in long format
ls -lR #lists the content of all entries in home directory

To remove a file, use the *rm* command. For example:

rm incoming/badfile

5.3.3.7 Debugging ascp Transfers

To make a test downloads using *ascp* please try this command:

ascp -i <private key file> -QTr <file(s) to transfer> -l100M asp- @upload.ncbi.nlm.nih.gov:test/

where

<private key file>	::= fully qualified path & file name where the generated private key was saved.
<files(s) to transfer>	::= names of files to transfer (including path)
	::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence.
100M	::= tunable mbit/sec bandwidth

Be sure that the local storage is fast enough to sustain this rate. We have seen problems with download if the target storage is on slow network volume. If you wish, examine unix */var/log/messages* for a fasp log file, and send that to Aspera support.

Note that when a submitter uses a wild card for submissions, 0 length files matching the shell expansion are created in the destination directory. These placeholders can be present for a time before the actual download takes place. Therefore, some buffer time should be added to any process on the transmission side that is responsible for determining whether the transfer succeeded.

A connection error like this one may be due to expiration of license key, or incorrect private key:

```
ascp: session open failed.
>> ascp: (remote) failed to initiate session, consult log.
>> Ssh error: SSH connection failure: 130.14.29.99:22 Server reported
>> failure exit code 1
```

5.3.3.8 Caveats

Supplying a directory as a source will cause the creation of the corresponding sub-directory tree on the destination. To avoid this, ensure that you execute the ascp command while in the source directory and provide a list of files to be transferred.

5.3.3.9 Known Problems

Please be aware that ':' (colon) character is not allowed in filenames by ascp command and files need to be renamed prior to transfer.

6 Tracking Submissions

The Sequence Read Archive Submissions page tracks submissions by SRA number and current status. There are two tabs: Completed Submissions, and Attention. Please look at the Attention tab to track any problems that might have arisen from submission. Please write to NCBI with any questions about why a submission has not completed.

7 Preparing Submissions

The best way to submit to the SRA is using the Interactive Submission Tool . You can prepare the metadata objects, and enter run filenames and checksums here. You must separately transmit data to NCBI using one of the abovementioned means.

7.1 Preparing Run Data

The SRA is intended as a repository of data output by "primary analysis" phase of the sequencing platform: sequencing results in fasta form along with instrument data indicating probability of correctness for each basecall (qualities) and signal intensity measurements (intensities).

The Sequence Read Archive does NOT accept fasta only datasets due to the inability to evaluate the quality of such data.

7.1.1 Roche/454

The SRA accepts deposits of sequencing read data from the 454 platform in the *.sff* format. These files should reflect the sequencing run setup. If the entire picotitre plate was used, then one *.sff* file per run should be submitted. If on the other hand the picotitre plate was divided into two or more regions, then a *.sff* file for each region should be submitted. If a *.sff* file contains more than one run, or more than one region in the run, please break up this file into constituent parts using the *sfffile* utility from the "Off Rig" software package provided by Roche.

Data Series	Number of Channels	Description
.sff	1	Flowgram (base call, phred quality score, flow value)

The read names found in the *.sff* file are meaningful and reflect the addressing scheme for the picotitre plate as well as a globally unique run id. Please do not rewrite this name as such addressing information will be lost. The sff file format is nearly optimal in terms of footprint, so there is little to be gained by further compressing them. Therefore, please provide *.sff* files uncompressed.

The sequencing data may have been produced by the 454 contract sequencing center (454MSC). Please ask 454MSC to provide *.sff* files for your project.

7.1.2 Illumina Genome Analyzer

7.1.2.1 Illumina native data

Original versions of the Illumina pipeline produced the following text files. These can be tarred together in packages comprising one lane's worth of data, and submitted as "Illumina_native" filetype. Please also obtain the flowcell name, lane id (1-8) for your run so that it can be used in submission.

Data Series	Number of Channels	Description
_seq.txt	1	Base calls per read
_prb.txt	4	Per channel log odds quality scores

7.1.2.2 Illumina SRF

For Illumina pipeline versions 1.1 and 1.3, the conduit for primary data is the sequence read format (SRF). Users should download the Staden io_lib package in order to get the *solexa2srf* utility. As an alternative to SRF, NCBI currently supports "native file" format submission[1]. These should be organized into a compressed tape archive file (*.tar.gz*), with all the files from each lane constituting one tar file.

To produce a primary analysis SRF submission file for a lane's worth of data, change the working directory to the run folder and do:

```
illumina2srf -R -P -N <run>:%l:%t: -n %x:%y
-o <center_name>_<run>_<lane>.srf s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run (for example `080117_EAS56_0068`), and <lane> is the desired lane.

To produce a primary analysis SRF submission file for a lane's worth of paired-ends data, change the working directory to the run folder and do:

```
illumina2srf -R -P -N <run>:%l:%t: -n %x:%y -2 <cycle>
-o <center_name>_<run>_<lane>.srf s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run, <lane> is the desired lane, and <cycle> indicates the cycle number that starts the second read.

Each flowcell contains 8 lanes but not all lanes are used for production. Also, some lanes are devoted to other projects. Finally, the size of the SRF file produced by this process

can be expected to be about 2 GB. For these reasons, it is desirable to produce one SRF file per lane. The SRF file format is nearly optimal in terms of footprint, so there is nothing to be gained by further compressing them. Therefore, please provide *.srf* files uncompressed.

7.1.2.3 Illumina Text Formats

In pipeline releases 1.4 and later, Illumina switched to text only forms of data. There are a variety of these forms depending on the pipeline version, the point at which data is extracted from the pipeline (pre or post-alignment), and whether certain features such as bar coding or paired ends are being used. Please contact NCBI if you plan to submit this kind of data.

7.1.3 Applied Biosystems SOLiD System

Primary analysis data from the SOLiD System is delivered in "color space", without translation into base space. Quality scores and signal intensities are based on the color calls.

7.1.3.1 SOLiD Native Format

NCBI currently supports "SOLID_native" format submission. These should be organized into a compressed tape archive file (*.tar.gz*), with all the files from one run constituting one tar file.

Sequencing data with minimal instrumentation output is appropriate for applications where the main goal is abundance measurement rather than reconstruction of original sequence.

Data Series	Number of Channels	Description
.csfasta	1	Base calls per read in color space
_QV.qual	1	Color space quality scores

For paired end data two files of each file type will exist (F3 and R3).

7.1.3.2 SOLiD SRF Format

NCBI recommends submission of SOLiD data in SRF format. Please download the SRF conversion utility at http://solidsoftwaretools.com/gf/project/srf/ .

7.1.4 Helicos HeliScope

NCBI is now taking datasets from the HeliScope. Please write us at trace@ncbi.nlm.nih.gov for special instructions.

8 Preparing Metadata

A salient feature of the SRA is the distinction given to metadata. Rather than embedding these with every run record, sequence read metadata are organized into a collection of XML files that capture as much, or as little, information as the submitter cares to give. Many pieces of information can be provided in the form of links and tag-value pairs, eliminating the need to negotiate complicated data representation ontologies.

Submission Object	Description	XML Schema specification
Study	XML file specifying sequencing study	SRA.study.xsd
Sample	XML file specifying the target of sequencing	SRA.sample.xsd
Experiment	XML file specifying experimental organization and parameters	SRA.experiment.xsd
Run	One of more XML descriptors linking run data to their experiments.	SRA.run.xsd

8.1.1 Genome Project Registration

Whole genome sequencing projects should be registered with the Entrez Genome Projects resource at NCBI before submitting to SRA. Please access the Entrez Genome Project Submission Form to submit.

8.1.2 Taxonomy Registration

Most single organism genome and transcriptome sequencing projects need a Taxon Id to help specify the sample being sequenced. Please consult the Entrez Taxonomy resource to see whether your organism is represented, and request an entry to be created if not. The taxon id is needed for submission preparation.

8.1.3 Reference Fields and Namespaces

All the XML files can take either names (aliases) to identify dependencies. These names need only be unique throughout the submission. Eventually, the SRA will replace these names with actual accessions while preserving the referential integrity of the records.

8.1.4 Required Fields

Each XML file has certain required fields. The XML schema document these entries and most are self-explanatory. Decisions that the submitter should make include:

Study type

Whether a genome project id exists for the study

The center project name or id

Whether a taxon id exists for the sample

Whether an anonymous id exists for the sample

Sequencing platform used (for example, 454 GS 20 or 454 GS FLX)

Library source, strategy, selection, layout if applicable, protocol

Spot or cluster layout (use of adapters, linkers, bar codes, etc)

Read processing selection

In addition, the submitter should think about the relationship between experiments and samples, runs and experiments, and whether to split any of these objects to represent distinct information

Finally, submitters may consider providing ancillary information, including links, Entrez links, and attribute tag-value pairs. These can be created for any of the five record types.

8.2 Preparing Submission Files

Aspects of the submission process pertaining to the submission itself have been broken out into their own XML descriptor. Contact information, transaction requests, exceptions, and file manifests can be listed here. Contacts should be provided for questions or problems pertaining to the particular submission.

Submission Object	Description	XML Schema specification
Submission	XML file specifying submission session	SRA.submission.xsd

A checksum should be computed for each run file delivered as part of the submission and entered into the *submission.xml* record. Please use the unix *md5sum* or equivalent utility. It is not necessary to provide checksums for the metadata xml files.

8.3 New Submission Protocol

Check your XML files for correctness with respect to the current published schema, for example:

```
xmllint --schema
'http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA/SR
A.run.xsd?view=co my.run.xml' > /dev/null
```

Check your XML files for completeness and referential integrity.

Verify checksums on run and analysis files.

Open *ftp* to the trace ftp site for your Center.

Change directory to *./short_reads*

Deposit the files using the ftp *put* or *mput* command. The files may be rolled into a single tar archive file.

Confirm receipt of the submission in the SRA Tracking Page. Once processed you will be able to download the submitted data through this page. Unfortunately, while the provisional SRA is in operation processing is manual and so we cannot guarantee any particular response time. Please let us know if you are under a tight publication deadline and we will try to accommodate your needs.

Please write to us at mailto:sra@ncbi.nlm.nih.gov with any questions about status or access.

9 Managing Existing Submissions

9.1 Update Submissions

Submitters can update their records through the interactive submission tool. It is recommended that you revisit your submission in order to annotate it with publication links and additional sample information. Once loaded, a run record cannot be updated (please write to NCBI if you wish to do this for some reason).

Bulk updates can be performed on the Project, Sample, and Experiment objects by submitting replacement XML files for the affected objects. An update XML document must identify the target accession and use the MODIFY action.

9.2 Hold Until Publish

An essential feature of the SRA is the ability to hold a submission until a manuscript reporting on the research is accepted or released by a journal:

Hold until date – This is appropriate for scheduled release of a publication

Release – The dataset can be released immediately to the public.

The hold can expire its term, or the submitter may send a Release message to NCBI indicating that the submission can be released to the public. Minimal tracking information about the submission including accession, date, center, title, platform, and size statistics are displayed on the SRA tracking page regardless of hold status.

A Release message can apply to the entire SRA object, or individual objects within the SRA submission. Any dependent objects are implicitly released. For example, releasing a certain experiment has the effect of releasing all its runs as well.

9.3 Versioning

SRA submissions are not explicitly versioned. Rather, a complete change history is stored for metadata and any version of the metadata can be accessed. Content such as run and analysis data are never modified. If these must be changed then current ones are deprecated (Withdrawn) and replacements added.

9.4 Curation

From time to time NCBI needs to update metadata in order to correct mistakes, propagate changes in other resources (for example taxonomic changes), and edit information in order to comply with editing requirements, copyrights, and data release policies. These "curation" changes may occur without necessarily seeking the approval of the original submitter. Run and analysis data will never be changed in this way. Also, original titles, descriptions, and names will be preserved as much as possible.

9.5 Suppression

At submitter's request, a certain record (submission, study, experiment, sample, run) can be suppressed.

Suppression simply marks the record as deprecated. Suppressed records are never actually deleted (except for technical reasons including for example a loading error). Suppressed records can still be accessed by accession, but the accession will be marked as having been suppressed. Suppressed records are not indexed, and copies of them are removed from download facilities.

Please write to NCBI if you wish to request suppression of a record or dataset.